Contact: Richard_Platt@harvard.edu

HIT Policy Committee & HIT Standards Committee
PCAST Workgroup

Richard Platt's invited remarks on the PCAST report's recommendations regarding Population Health, including comments on distributed data networks.

These comments focus principally on the PCAST report's chapter VII, "Health Data and the Research Opportunity," which addresses the benefits of improved access to real-time, real-world, comprehensive data. Specific examples mentioned include: syndromic surveillance and public health monitoring, monitoring of adverse events associated with medical products, assessment of dissemination and utilization (including quality assessment), and comparative effectiveness research. I base these remarks on my experience and that of my colleagues in these domains, in partnership with the Centers for Disease Control, the Food and Drug Administration, the Agency for Healthcare Research and Quality, and the Massachusetts Department of Public Health.

The PCAST recommendations appropriately focus on the use of data to support the delivery of medical care. They are not, however, not well suited for secondary uses, such as the ones I described above. For example, the report recommends that deidentifed study-specific datasets be created on an as needed basis, containing information from individuals who have agreed to have their data used in this way. The datasets will be provided to public health agencies or researchers for analysis.

I believe this approach will be problematic for five reasons:

1. Logistical complexity,
2. Threats to privacy,
3. Threats to validity caused by a requirement for individual consent to use the data,
4. Risk of misinterpretation of results,
5. Reluctance by data holders to provide the data.

As an example of the kind of question we would want to address, consider the FDA Mini-Sentinel program's planned evaluation of the impact of different diabetes treatments on the risk of myocardial infarction (1). This evaluation requires repeated assessment of over a million people with diabetes, some of whom initiate therapy with one of five different treatment regimens. Every 3 months it will be necessary to:

1. Identify all new users of one the 5 treatments (includes affirmative evidence of non-use in the preceding year).
2. Restrict the analysis to individuals who meet all the following criteria:
    a. Continuous prescription drug coverage for the preceding 12 months
    b. No hospital discharge within the 60 days preceding the first prescription with a principal diagnosis of acute myocardial infarction
    c. No prior use of the new treatment identified in step 1 during the preceding 12 months
    d. At least one other dispensing of a diabetes drug within the preceding year OR at least one diagnosis of diabetes

3. Determine eligible person time, that is, the intervals for each person during which we can be confident that both drug exposures and health outcomes are captured (omitting periods with no health insurance, or no prescription drug coverage, for instance).
4. Obtain information on all diagnoses, all procedures, and all other medications from the year before initiation through the end of followup. Information on non-diabetes treatments and diagnoses are needed in order to adjust for the baseline risk of myocardial infarction.
5. Update information for individuals who were identified previously.
6. Check the information for completeness and quality.
7. Compute risk differences between individuals exposed to different regimens, adjusted for the individuals' baseline risk of myocardial infarction.

Logistics

The PCAST report doesn't describe how such a process will work. However, it appears to require repeatedly creating centralized datasets containing vast amounts of sensitive information. Simply moving, storing, and analyzing that data will impose serious logistical burdens. Each time new data is added, the evaluation team will need either to reconstruct the entire database, or have sufficient identifying information to permit merging individuals' new data to the information that had been transmitted previously. This will require the organizations that provide the data to maintain and use separate lists for every study to match new data to the data for previously deidentified records. In practice, it is complex, time consuming, and expensive to accomplish this. This problem will be compounded by the need to create new, standalone datasets for every question. For example, a dataset created to study the safety of a diabetes drug will not be useful to study an arthritis drug, even though there will be substantial overlap in the populations.

Privacy

I am also concerned that it will be difficult, under the suggested paradigm, to maintain the confidentiality of personal health information. Many of the intended secondary uses of EHR and claims data depend on access to nearly complete health information covering several years for every person. The completeness of the information allows many possibilities for identifying the individual, even if direct identifiers like name and address are removed. For example, knowing only the date and hospital of an individual's surgical procedure can often allow identification of that person in the dataset with reasonably high precision. It would then be possible to identify all of the other information about that person that is contained in the dataset. This could include a complete medication history and all diagnoses, including ones unrelated to the surgical procedure.

Consent

For several of the intended purposes, it is either desirable or necessary to include the entire population of a geographic region (syndromic surveillance), or everyone exposed to a new drug (safety monitoring), or all the patients cared for by a health care system (quality). Even when completeness is not necessary, it is essential to evaluate a representative population. This means that a requirement to obtain individual consent would make the results uninterpretable. There is little evidence to support the PCAST report's assertion that it is possible to adjust fully for the

biases that could be introduced by requiring consent. Many people believe the claim is incorrect, or, at best, not provable.

In current practice, public health uses of confidential medical information without individual consent are addressed in the Health Insurance Portability and Accountability Act (HIPAA) Research uses of data are governed by HIPAA and other regulations. Institutional review boards, which oversee research involving human subjects, are specifically authorized to waive individual consent under certain circumstances. It will be important to preserve these established mechanisms for approving and overseeing the use of electronic health information without individual consent in implementing the PCAST report's recommendations.

Interpretation of results

The proposed approach will cause problems in accurately interpreting data from different sources. Both EHR and claims data are often customized to support local needs and preferences. One example is differences in the specific diagnoses that are assigned for the same condition, resulting from the fact that clinicians customize lists of frequently used diagnoses in EHRs. Identifying, understanding, and addressing these differences can require the active participation of local experts. This kind of engagement is difficult to perform after the data from different organizations has been centralized.

Willingness to provide data

Finally, the requirement to transfer the large amounts of data that are required reduces the willingness of data holders to make their data available. This is because hospitals, medical practices, and payors understand that the same data can also be used for other purposes, for instance by competitors seeking market advantage.

Distributed data, distributed analysis networks as an alternative

An alternative is to use distributed data networks that allow questions to be answered without requiring data to be centralized for analysis (2, 3). The basic premise of these distributed networks is that much or all of the required data manipulation can be performed by the covered entities or other organizations that already possess the information. These entities can then share the results – information – rather than data.

Examples of such networks include the CDC's Vaccine Safety Datalink (4), the Observational Medical Outcomes Partnership (5), the FDA's Mini-Sentinel program (6), and the National Bioterrorism Syndromic Surveillance Demonstration Program (7).

Each of these distributed networks has employed the same basic approach:

1. They established a common data model that specified the names, formats, and definitions of data elements of interest. These data models have not needed to capture all of the complexity in the underlying data to address the questions of interest.
2. Each site created a copy of the relevant data that had been transformed into the format of the common data model.
3. Each site kept the transformed data behind its firewall,
4. The evaluators distributed computer programs to perform desired analyses at the sites. Most analyses require many programs to be distributed and executed in sequence. Some

of these check the completeness and quality of the data. Some programs provide descriptive information about the population, its health status, and its treatment and outcomes. Other programs perform statistical analyses.

5. The sites returned the programs' output, which typically consisted of highly aggregated information such as counts, rates, or adjusted rates,
6. The evaluators merged the results from the separate sites to yield a result that reflected all of the information.

Distributed data, distributed analysis networks have shown that they can address important questions while minimizing the amount of data that must be provided to third parties, avoiding disclosure of potentially identifying information, facilitating engagement of individuals who understand the data, and providing assurance to data holders that the data will not be used for other purposes.

In some situations, it is necessary to create a pooled person-level database. In this setting, the information needed is typically a small fraction of the total amount that contributes to the full analysis. For instance, the shared information might state that the age is 60 to 65 years old, rather than providing an exact age or birthdate. It might also state the number of days between first exposure to a drug and a subsequent myocardial infarction, without providing either the date of exposure or the date of the hospitalization. And it is possible to provide a computed illness severity score, rather than all the diagnoses, procedures and treatments that contribute to the score.

Conclusion

In summary, large amounts of data for millions of people are required to address many critical questions regarding the effectiveness, safety, and quality of medical care, and to support other essential societal needs. The distributed-data, distributed-analysis framework can often accomplish this in ways that avoid the problems that will be encountered in using the PCAST report's approach.

1. www.mini-sentinel.org/work_products/Evaluations/AMI_Surveillance_Protocol_and_Appendices_ABC.pdf

2: Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010;48(6 Suppl):S45-51.

3 Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Ann Intern Med. 2009 Sep 1;151(5):341-4.

4. Lieu TA, Kulldorff M, Davis RL, Lewis EM, Weintraub E, Yih K, Yin R, Brown JS, Platt R; for the Vaccine Safety Datalink Rapid Cycle Analysis Team. Real-time vaccine safety surveillance for the early detection of adverse events. Med Care. 2007;45 (10 Supl 2):S89-95.

Contact: Richard_Platt@harvard.edu

5: Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153:600-6.

6. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System - A National Resource for Evidence Development. N Engl J Med. 2011; 364:498-499.

7: Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. BMC Public Health. 2006;6:235.